BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35190
doi: 10.3897/biss.3.35190

OPEN ACCESS

Conference Abstract

A Workflow for Data Extraction from Digitized Herbarium Specimens

Sohaib Younis[‡,§], Marco Schmidt[‡,|], Bernhard Seeger[§], Thomas Hickler[‡,¶], Claus Weiland[‡]

‡ Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany
§ Department of Mathematics and Computer Science, Philipps University, Marburg, Germany
| Palmengarten, Frankfurt am Main, Germany
¶ Goethe University, Frankfurt am Main, Germany

Corresponding author: Sohaib Younis (muhammad-sohaib.younis@senckenberg.de)

Received: 05 Apr 2019 | Published: 10 Jul 2019

Citation: Younis S, Schmidt M, Seeger B, Hickler T, Weiland C (2019) A Workflow for Data Extraction from Digitized Herbarium Specimens. Biodiversity Information Science and Standards 3: e35190. https://doi.org/10.3897/biss.3.35190

Abstract

Based on own work on species and trait recognition and complementary studies from other working groups, we present a workflow for data extraction from digitized herbarium specimens using convolutional neural networks. Digitized herbarium sheets contain:

1. preserved plant material as well as additional objects:
2. the label containing information on the collection event,
3. annotations such as revision labels, or notes on material extraction,
4. identifiers such as barcodes or numbers,
5. envelopes for loose plant material and
6. often scale bars and color charts used in the digitization process.

In order to treat these objects appropriately, segmentation techniques (Triki et al. 2018) will be applied to localize and identify the different kinds of objects for specific treatments. Detecting presence of plant organs such as leaves, flowers or fruits is already a first step in data extraction potentially useful for phenological studies. Plant organs will be subject to routines for quantitative (Gaikwad et al. 2018) and qualitative (Younis et al. 2018) trait recognition routines. Text-based objects can be treated as described by Kirchhoff et al. 2018, using OCR techniques and considering the many collection-specific terms and

© Younis S et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

abbreviations as described in Schröder 2019. Additionally, species recognition (Younis et al. 2018) will be applied in order to help further identification of incompletely identified collection items or to detect possible misidentifications. All steps described above need sufficient training data including labelling that may be obtained from collection metadata and trait databases.

In order to deal with new incoming digitized collections, unseen data or categories, we propose implementation of a new Deep Learning approach, so-called Lifelong Learning: Past knowledge of the network is dynamically saved in latent space using autoencoder and generatively replayed while the network is trained on new tasks which enables it to solve complex image processing tasks without forgetting former knowledge while incrementally learning new classes and knowledge.

Keywords

Trait Recognition, Convolutional Neural Networks, Lifelong Learning, Herbarium specimens, Trait Semantics, Digitized Natural History Collections, Image Processing, Image Captioning, Object Detection, Object Annotation

Presenting author

Sohaib Younis

References

- Gaikwad J, Triki A, Bouaziz B, Hamed H, Hentschel J (2018) TraitEx: tool for measuring morphological functional traits from digitized herbarium specimens. Proceedings of the 10th International Conference on Ecological Informatics. Jena
- Kirchhoff A, Bügel U, Santamaria E, Reimeier F, Röpert D, Tebbje A, Güntsch A, Chaves F, Steinke K, Berendsohn W (2018) Toward a service-based workflow for automated information extraction from herbarium specimens. Database 2018 https://doi.org/10.1093/database/bay103
- Schröder CN (2019) Katalog der auf Herbarbelegen gebräuchlichen Abkürzungen. Kochia 12: 37-67.
- Triki A, Bouaziz B, Gaikwad J (2018) Refined methodology for accurately detecting objects from digitized herbarium specimens. Proceedings of the 10th International Conference on Ecological Informatics. Jena
- Younis S, Weiland C, Hoehndorf R, Dressler S, Hickler T, Seeger B, Schmidt M (2018) Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. Botany Letters 165: 377-383. https://doi.org/10.1080/23818107.2018.1446357